3

Multivariate Models and Algorithms for Learning Correlation Structures from Replicated Molecular Profiling Data

Lipi R. Acharya[1] and Dongxiao Zhu[1,2]
[1]*University of New Orleans, New Orleans*
[2]*Research Institute for Children, Children's Hospital, New Orleans*
U.S.A.

1. Introduction

Advances in high-throughput data acquisition technologies, e.g. microarray and next-generation sequencing, have resulted in the production of a myriad amount of molecular profiling data. Consequently, there has been an increasing interest in the development of computational methods to uncover gene association patterns underlying such data, e.g. gene clustering (Medvedovic & Sivaganesan, 2002; Medvedovic *et al.*, 2004), inference of gene association networks (Altay and Emmert-Streib, 2010; Butte & Kohane, 2000; Zhu *et al.*, 2005), sample classification (Yeung & Bumgarner, 2005) and detection of differentially expressed genes (Sartor *et al.*, 2006). However, outcome of any bioinformatics analysis is directly influenced by the quality of molecular profiling data, which are often contaminated with excessive noise. Replication is a frequently used strategy to account for the noise introduced at various stages of a biomedical experiment and to achieve a reliable discovery of the underlying biomolecular activities.

Particularly, estimation of the correlation structure of a gene set arises naturally in many pattern analyses of replicated molecular profiling data. In both supervised and unsupervised learning, performance of various data analysis methods, e.g. linear and quadratic discriminate analysis (Hastie *et al.*, 2009), correlation-based hierarchial clustering (Eisen *et al.*, 1998; de Hoon *et al.*, 2004; Yeung *et al.*, 2003) and co-expression networking (Basso *et al.*, 2005; Boscolo *et al.*, 2008) relies on an accurate estimate of the true correlation structure.

The existing MLE (maximum likelihood estimate) based approaches to the estimation of correlation structure do not automatically accommodate replicated measurements. Often, an *ad hoc* step of data preprocessing by averaging (either weighted, unweighted or something in between) is used to reduce the multivariate structure of replicated data into bivariate one (Hughes *et al.*, 2000; Yao *et al.*, 2008; Yeung *et al.*, 2003). Averaging is not completely satisfactory as it creates a strong bias while reducing the variance among replicates with diverse magnitudes. Moreover, averaging may lead to a significant amount of information loss, e.g. it may wipe out important patterns of small magnitudes or cancel out opposite patterns of similar magnitudes. Thus, it is necessary to design multivariate correlation estimators by treating each replicate exclusively as a random variable. In general, the experimental design that specifies replication mechanism of a gene set may be unknown

(blind) or known (informed) to data analysts. The suite of multivariate models and algorithms offer flexible ways to capture the correlation structure of a gene set with diverse replication mechanisms and allow for further generalizations.

In this chapter, we present bivariate and multivariate approaches to estimate the correlation structure of a gene set with replicated measurements. We begin with two popular bivariate correlation estimators, Pearson's correlation (Eisen *et al.*, 1998; Kung *et al.*, 2005) and SD-weighted correlation (Hughes *et al.*, 2000; Yeung *et al.*, 2003) followed by a comprehensive discussion of three generalized multivariate models, blind-case model, informed-case model and finite mixture model introduced in (Acharya & Zhu, 2009; Zhu *et al.*, 2007; 2010) to estimate the correlation structure of a gene set with either blind or informed replication mechanism. We analyze the performance of various correlation estimators using synthetic and real-world replicated data sets.

2. Replicated molecular profiling data

Molecular profiling data in the present context refers to a numerical matrix of gene abundance levels, where rows correspond to genes and columns represent experiments (samples). High-throughput platforms, such as microarrays, enable the scientists to simultaneously interrogate the expression abundance of tens of thousands of genes in the living cell. A microarray experiment is typically performed by hybridizing target cRNA samples labeled with fluorescent dyes on a glass slide spotted with oligonucleotides. After hybridization, the glass slide is washed and scanned to detect the gene expression levels. Some of the popular microarray platforms include Affymetrix GeneChip, Agilent Microarray, Illumina BeadArray and housemade twocolor arrays. Based on the experimental design employed by a data acquisition platform, the replication mechanism underlying molecular profiling data can be either *blind* or *informed* to data analysts (Figure 1). For example, the measurements from Affymetrix GeneChip platform (Lokhart *et al.*, 1996) correspond to blind replication mechanism, where expression levels of a gene are measured by designing a set of 11 perfect match sibling probes against the 3-prime end of mRNA, although a mixture of gene isoforms can exist. On the other hand, some of the more recent Illumina hybridization-based BeadArray (Gunderson *et al.*, 2004) and deep sequencing based Genome Analyzer II (Shendure & Ji, 2008) platforms utilize an informed replication mechanism. Indeed, such platforms simultaneously profile $6 - 12$ samples of whole-genome gene expression in a chip, where both biological and technical replicates can be used in the experiment. Many studies also use a more general replication strategy of combining the two mechanisms, e.g. blind replication mechanism nested within the informed mechanism and *vice versa* (Kerr & Churchill, 2001). It is necessary to explicitly consider both blind and informed mechanisms for a robust pattern analyses of replicated data. For instance, Fig. 1 presents two gene sets with the same number of replicated measurements, however, their underlying correlation structures differ by incorporating the prior knowledge of replication mechanism. For a comprehensive correlation based analysis of replicated molecular profiling data with both blind and informed replication mechanism, we refer to (Zhu *et al.*, 2010).

3. Bivariate correlation estimators

In this section, we discuss two bivariate correlation estimators, Pearson's correlation (Eisen *et al.*, 1998; Kung *et al.*, 2005; Rengarajan *et al.*, 2005) and SD-weighted correlation (Hughes

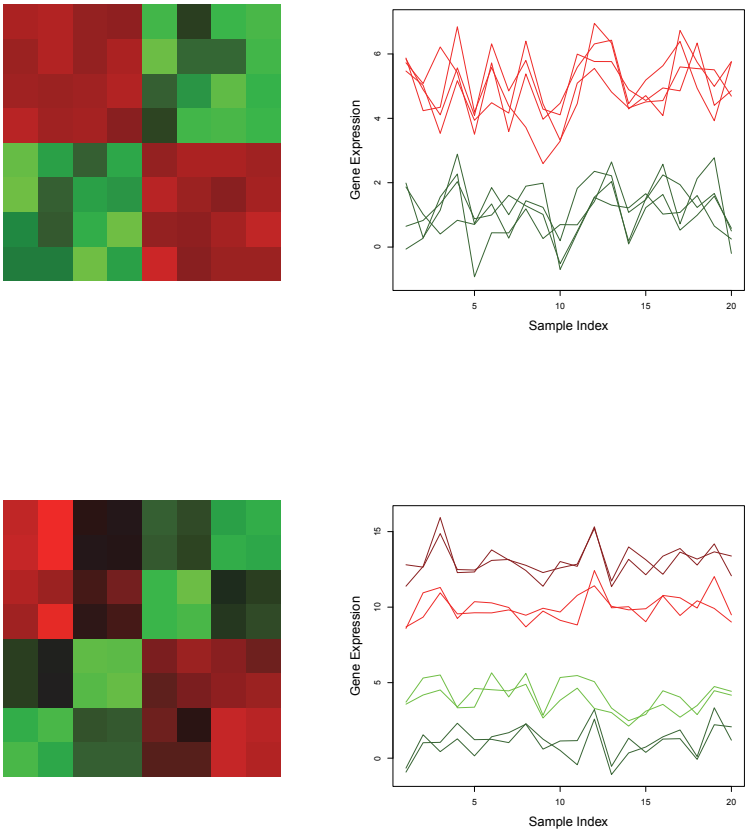

Fig. 1. Correlation structures (left) and molecular profiling data (right) corresponding to a pair of genes, each with 4 replicated measurements. The upper panels represent the correlation structure and molecular profiling data with blind replication mechanism, whereas the lower panels correspond to the ones with informed replication mechanism. In case of informed replication mechanism 2 biological replicate and 2 technical replicates nested within each biological replicates are used for a gene.

et al., 2000; van't Veer *et al.*, 2002; Yeung *et al.*, 2003), frequently used in the analysis of replicated molecular profiling data. We assume that the abundance levels of two genes X and Y with m_1 and m_2 replicated measurements respectively, are simultaneously measured over n independent experiments. If x_{ij} and y_{ij} denote the abundance levels of X and Y in the i^{th} replicate and j^{th} sample respectively, we write

$$\bar{x}_j = \frac{1}{m_1} \sum_{i=1}^{m_1} x_{ij} \tag{1}$$

and

$$\bar{y}_j = \frac{1}{m_2} \sum_{i=1}^{m_2} y_{ij} \tag{2}$$

for the average measurements in the j^{th} sample,

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} \bar{x}_j \tag{3}$$

and

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} \bar{y}_j \tag{4}$$

for the grand means of the measurements,

$$s_x^2(j) = \frac{1}{m_1 - 1} \sum_{i=1}^{m_1} (x_{ij} - \bar{x}_j)^2 \tag{5}$$

and

$$s_y^2(j) = \frac{1}{m_2 - 1} \sum_{i=1}^{m_2} (y_{ij} - \bar{y}_j)^2 \tag{6}$$

for the variances in the j^{th} sample,

$$\bar{x}_w = \sum_{j=1}^{n} \frac{\bar{x}_j}{s_x^2(j)} \Big/ \sum_{j=1}^{n} \frac{1}{s_x^2(j)} \tag{7}$$

and

$$\bar{y}_w = \sum_{j=1}^{n} \frac{\bar{y}_j}{s_y^2(j)} \Big/ \sum_{j=1}^{n} \frac{1}{s_y^2(j)}, \tag{8}$$

for the SD-weighted average measurements corresponding to X and Y, $j = 1, \ldots, n$.

3.1 Pearson's correlation estimator

Pearson's correlation coefficient is a well-known similarity measure for clustering molecular profiling data (Eisen *et al.*, 1998). The estimate of correlation between X and Y is defined in terms of unweighted average of replicated measurements for a gene across different experiments (Kung *et al.*, 2005; Rengarajan *et al.*, 2005) and is given by

$$cor(X, Y) = \frac{\sum_{j=1}^{n} (\bar{x}_j - \bar{x})(\bar{y}_j - \bar{y})}{\sqrt{\sum_{j=1}^{n} (\bar{x}_j - \bar{x})^2 \sum_{j=1}^{n} (\bar{y}_j - \bar{y})^2}}. \tag{9}$$

In case of a gene set with k genes $X_1, \ldots, X_k$, where m_i replicated measurements are available for X_i, the correlation structure is defined by all pairwise correlations $cor(X_i, X_j)$, $i, j = 1, \ldots, k$. Due to its closed-form representation, Pearson's estimator enjoys computational simplicity. However, it is exclusively based on estimating bivariate correlation from a data with multivariate structure. Additionally, the estimator assigns equal weights to all replicates of a gene without considering the variation in their magnitudes, which is often large for data generated from high-throughput platforms. To overcome this problem, a number of more generalized correlation estimators have been proposed by considering weighted average of replicated measurements in place of simple average.

3.2 SD-weighted correlation estimator

The SD-weighted correlation estimator considers weighted average of replicated measurements, where weights are determined by standard deviations of the measurements across different experiments. The SD-weighted correlation between X and Y is defined as (Hughes *et al.*, 2000; Zhu *et al.*, 2010)

$$cor_w(X, Y) = \frac{\sum_{j=1}^{n} \left(\frac{\bar{x}_j - \bar{x}_w}{s_x(j)} \right) \left(\frac{\bar{y}_j - \bar{y}_w}{s_y(j)} \right)}{\sqrt{\sum_{j=1}^{n} \left(\frac{\bar{x}_j - \bar{x}_w}{s_x(j)} \right)^2 \sum_{j=1}^{n} \left(\frac{\bar{y}_j - \bar{y}_w}{s_y(j)} \right)^2}}. \tag{10}$$

Advantages of SD-weighted correlation have been demonstrated in terms of increased accuracy and stability in cluster analysis, compared with Pearson's estimator (Yeung *et al.*, 2003). Nevertheless, SD-weighted estimator also does not explicitly accommodate replicated measurements and requires a preprocessing of data by computing their weighted average. In averaging, many useful patterns of small magnitude may be wiped out or patterns of opposite magnitude may be canceled out. Moreover, standard deviation of replicated measurements may not be a faithful representation of their internal variation, specially when the number of replicates is small. This problem has been addressed by considering a shrinkage version of the correlation estimator (Yao *et al.*, 2008), however, none of the aforementioned estimators are ready to explicitly accommodate replicated data and exploit prior knowledge of experimental design that explains replication mechanism.

4. Multivariate correlation estimators

In this section, we review three multivariate models, blind-case model (Acharya & Zhu, 2009; Zhu *et al.*, 2007), informed-case model (Zhu *et al.*, 2010) and finite mixture model (Acharya & Zhu, 2009) for estimating the correlation structure from replicated measurements corresponding to a gene set with blind or informed replication mechanism. Throughout this section, we treat each replicated measurement individually as a random variable and assume that data are independently and identically distributed samples from a multivariate normal distribution. We discuss the parameter structures for each model and their estimation from replicated measurements corresponding to a pair of genes X and Y or a gene set with k genes $X_1, \ldots, X_k$. It is assumed that gene abundance levels are measured over n independent samples, where m_i replicated measurements of the i^{th} gene X_i are available in each of them, $i = 1, \ldots, k$. We denote the n multivariate samples by Z_j, $j = 1, \ldots, n$.

4.1 Blind-case model

Blind-case model from (Acharya & Zhu, 2009; Zhu *et al.*, 2007) estimates the correlation structure of a gene set with replicated measurements by assuming a constrained set of parameters in the multivariate normal distribution. The model is designated as 'blind' since it imposes a fixed number of within-molecular and between-molecular correlation parameters in the underlying correlation structure. Throughout this section, we follow the notations from (Acharya & Zhu, 2009). The parameters, mean vector μ_B and the correlation matrix Σ_B, for the blind-case model are defined as

$$\mu_B = \begin{bmatrix} \mu^B_{x_1} e_{m_1} \\ \vdots \\ \mu^B_{x_k} e_{m_k} \end{bmatrix} \tag{11}$$

where $\mu^B_{x_i}$ is a scalar and $e_{m_i} = (1,\ldots,1)^T$ is a vector of size $m_i \times 1$, for $i = 1,\ldots,k$. The correlation matrix Σ_B of size $\sum_{i=1}^k m_i \times \sum_{i=1}^k m_i$ has the following structure

$$\begin{aligned} \Sigma_B &= \begin{bmatrix} 1 & \cdots & \rho_{11} & \cdots & \rho_{1k} & \cdots & \rho_{1k} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ \rho_{11} & \cdots & 1 & \cdots & \rho_{1k} & \cdots & \rho_{1k} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ \rho_{k1} & \cdots & \rho_{k1} & \cdots & 1 & \cdots & \rho_{kk} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ \rho_{k1} & \cdots & \rho_{k1} & \cdots & \rho_{kk} & \cdots & 1 \end{bmatrix} \\ &= \begin{bmatrix} \Sigma^B_{11} & \cdots & \Sigma^B_{1k} \\ \vdots & \vdots & \vdots \\ {\Sigma^B_{1k}}^T & \cdots & \Sigma^B_{kk} \end{bmatrix}, \end{aligned} \tag{12}$$

where Σ^B_{ij} is a $m_i \times m_j$ submatrix defined in terms of a single parameter ρ_{ij}. The parameters ρ_{ij}'s correspond to either within-molecular correlation (case $i = j$) or between-molecular correlation (case $i \neq j$). As a correlation matrix is symmetric, it is assumed that $\rho_{ij} = \rho_{ji}$. For practical purposes, only between-molecular correlations are of interest, whereas within-molecular correlations indicate data quality. Indeed, higher values of within-molecular correlations correspond to cleaner data.

To estimate the model parameters, the path of maximum likelihood estimation is followed. Due to their asymptotic properties, the MLE's are frequently used in parameter estimation problems when the underlying distribution is multivariate normal (Casella & Berger, 1990). Suppose the n observations Z_j's are sampled from multivariate normal distribution $N(\mu,\Sigma)$ with parameters μ and Σ, where $n > \sum_{i=1}^k m_i$. Then the likelihood function is defined as

$$L(\mu,\Sigma) = \prod_{j=1}^n N(Z_j|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{1}{2}(\sum_{i=1}^k m_i)n}|\Sigma|^{\frac{1}{2}n}} \exp[-\frac{1}{2}\sum_{j=1}^n (Z_j-\mu)^T\Sigma^{-1}(Z_j-\mu)]. \tag{13}$$

The MLE's are estimated by maximizing L with respect to μ and Σ. In the present context, if the the abundance level of l^{th} gene in its i^{th} replicate and j^{th} sample is denoted by x_{ij}^{l}, the MLE's of μ_B and Σ_B are obtained by solving

$$d\mathcal{L}/d\mu_{x_l}^{B} = 0, \tag{14}$$

for $l = 1, \ldots, k$ and

$$d\mathcal{L}/d\Sigma = 0, \tag{15}$$

where $\mathcal{L} = \log L$. This results in

$$\hat{\mu}_{x_l}^{B} = \frac{1}{n}\frac{1}{m_l}\sum_{j=1}^{n}\sum_{i=1}^{m_l} x_{ij}^{l} \tag{16}$$

for $l = 1, \ldots, k$. Thus, the MLE of μ_B is

$$\hat{\mu}_B = \begin{bmatrix} \hat{\mu}_{x_1}^{B} e_{m_1} \\ \vdots \\ \hat{\mu}_{k}^{B} e_{m_k} \end{bmatrix}. \tag{17}$$

The MLE of Σ_B is given by

$$\hat{\Sigma}_B = \frac{1}{n}\sum_{j=1}^{n}(Z_j - \hat{\mu}_B)(Z_j - \hat{\mu}_B)^T. \tag{18}$$

As the parameters $\hat{\rho}_{ij}$'s may not be tractable in practice, they are estimated using

$$\hat{\rho}_{ij} = Avg(\hat{\Sigma}_{ij}^{B}),\ i, j = 1, \ldots, k. \tag{19}$$

Equations 17-19 are used to obtain the correlation structure from blind-case model. When $k = 2$, blind-case model is defined in terms of two within-molecular and one between molecular correlation parameters, as presented in (Zhu *et al.*, 2007). Further, if there are no replicates for X and Y or $m_1 = m_2 = 1$, blind-case model and Pearson's correlation coefficient (Eq. 9) are connected as follows (Zhu *et al.*, 2007)

$$\hat{\rho}_{12} = \frac{n-1}{n} cor(X, Y). \tag{20}$$

Overall, blind-case model presents a simple and parsimonious multivariate approach for estimating the correlation structure of a gene set with blind replication mechanism. As the MLE's of parameters have closed-form representation, the model is computationally very efficient, e.g. it is well known that the infinite Bayesian mixture model approach (Medvedovic & Sivaganesan, 2002; Medvedovic *et al.*, 2004) suffers from non-trivial computational complexity as the number of genes and replicated measurements increases. However, blind-case model always imposes a fixed number of parameters in the model. This may correspond to an oversimplified representation of the underlying correlation structure of a gene set or an overly constrained correlation structure in case of replicated data for which the underlying experimental design is known. Thus, it is desirable to consider more flexible

multivariate models by explicitly incorporating prior knowledge of replication mechanisms in the correlation structure.

4.2 Informed-case model

Informed-case model introduced in (Zhu *et al.*, 2010) generalizes blind-case model by accommodating prior knowledge of replication mechanism. In many cases the number of biological and technical replicates used in the experimental design are known. Informed-case model utilizes this information and assigns different parameters for the biological replicates of a gene. For simplicity, we present the informed-case model for two genes X and Y, where 3 biological replicates and 2 technical replicates nested within each biological replicate are used for each of them. This representation can be naturally extended to the case of a gene set with a given number of biological and technical replicates. Throughout this section, we follow the notations from (Zhu *et al.*, 2010). The two parameters, mean vector μ_I and correlation matrix Σ_I, for the informed-case model are defined as

$$\mu^I = \left(\mu_x^1, \mu_x^1, \mu_x^2, \mu_x^2, \mu_x^3, \mu_x^3, \mu_y^1, \mu_y^1, \mu_y^2, \mu_y^2, \mu_y^3, \mu_y^3\right)^T \tag{21}$$

and

$$\Sigma^I = \begin{pmatrix}
1 & \rho^{tt} & \rho_x^{12} & \rho_x^{12} & \rho_x^{13} & \rho_x^{13} & \rho_{xy}^{11} & \rho_{xy}^{11} & \rho_{xy}^{12} & \rho_{xy}^{12} & \rho_{xy}^{13} & \rho_{xy}^{13} \\
\rho^{tt} & 1 & \rho_x^{12} & \rho_x^{12} & \rho_x^{13} & \rho_x^{13} & \rho_{xy}^{11} & \rho_{xy}^{11} & \rho_{xy}^{12} & \rho_{xy}^{12} & \rho_{xy}^{13} & \rho_{xy}^{13} \\
\rho_x^{21} & \rho_x^{21} & 1 & \rho^{tt} & \rho_x^{23} & \rho_x^{23} & \rho_{xy}^{21} & \rho_{xy}^{21} & \rho_{xy}^{22} & \rho_{xy}^{22} & \rho_{xy}^{23} & \rho_{xy}^{23} \\
\rho_x^{21} & \rho_x^{21} & \rho^{tt} & 1 & \rho_x^{23} & \rho_x^{23} & \rho_{xy}^{21} & \rho_{xy}^{21} & \rho_{xy}^{22} & \rho_{xy}^{22} & \rho_{xy}^{23} & \rho_{xy}^{23} \\
\rho_x^{31} & \rho_x^{31} & \rho_x^{32} & \rho_x^{32} & 1 & \rho^{tt} & \rho_{xy}^{31} & \rho_{xy}^{31} & \rho_{xy}^{32} & \rho_{xy}^{32} & \rho_{xy}^{33} & \rho_{xy}^{33} \\
\rho_x^{31} & \rho_x^{31} & \rho_x^{32} & \rho_x^{32} & \rho^{tt} & 1 & \rho_{xy}^{31} & \rho_{xy}^{31} & \rho_{xy}^{32} & \rho_{xy}^{32} & \rho_{xy}^{33} & \rho_{xy}^{33} \\
\rho_{xy}^{11} & \rho_{xy}^{11} & \rho_{xy}^{21} & \rho_{xy}^{21} & \rho_{xy}^{31} & \rho_{xy}^{31} & 1 & \rho^{tt} & \rho_y^{12} & \rho_y^{12} & \rho_y^{13} & \rho_y^{12} \\
\rho_{xy}^{11} & \rho_{xy}^{11} & \rho_{xy}^{21} & \rho_{xy}^{21} & \rho_{xy}^{31} & \rho_{xy}^{31} & \rho^{tt} & 1 & \rho_y^{12} & \rho_y^{12} & \rho_y^{13} & \rho_y^{12} \\
\rho_{xy}^{12} & \rho_{xy}^{12} & \rho_{xy}^{22} & \rho_{xy}^{22} & \rho_{xy}^{32} & \rho_{xy}^{32} & \rho_y^{21} & \rho_y^{21} & 1 & \rho^{tt} & \rho_y^{23} & \rho_y^{23} \\
\rho_{xy}^{12} & \rho_{xy}^{12} & \rho_{xy}^{22} & \rho_{xy}^{22} & \rho_{xy}^{32} & \rho_{xy}^{32} & \rho_y^{21} & \rho_y^{21} & \rho^{tt} & 1 & \rho_y^{23} & \rho_y^{23} \\
\rho_{xy}^{13} & \rho_{xy}^{13} & \rho_{xy}^{23} & \rho_{xy}^{23} & \rho_{xy}^{33} & \rho_{xy}^{33} & \rho_y^{31} & \rho_y^{31} & \rho_y^{32} & \rho_y^{32} & 1 & \rho^{tt} \\
\rho_{xy}^{13} & \rho_{xy}^{13} & \rho_{xy}^{23} & \rho_{xy}^{23} & \rho_{xy}^{33} & \rho_{xy}^{33} & \rho_y^{31} & \rho_y^{31} & \rho_y^{32} & \rho_y^{32} & \rho^{tt} & 1
\end{pmatrix}, \tag{22}$$

where ρ_x^{ij}, ρ_y^{ij} and ρ_{xy}^{ij} denote within-molecular and between-molecular correlations between i^{th} and j^{th} biological replicates. As the technical replicates of a biological replicate are often highly correlated, we use a single parameter ρ^{tt} to represent their correlation.

Analogous to the case of blind-case model (Eq. 14 and Eq. 15), the MLE's $\hat{\mu}^I$ and $\hat{\Sigma}^I$ are given by the following sets of equations

$$\hat{\mu}_x^{j_{m_1}} = \frac{1}{I_{m_1}^j n} \sum_{k=1}^{n} \sum_{i=\sum_{l=1}^{j} I_{m_1}^{l-1}+1}^{\sum_{l=1}^{j} I_{m_1}^{l}} x_{ik},\ 1 \le j_{m_1} \le J_{m_1} \tag{23}$$

$$\hat{\mu}_y^{j_{m_2}} = \frac{1}{I_{m_2}^j n} \sum_{k=1}^{n} \sum_{i=\sum_{l=1}^{j} I_{m_2}^{l-1}+1}^{\sum_{l=1}^{j} I_{m_2}^{l}} y_{ik},\ 1 \le j_{m_2} \le J_{m_2} \tag{24}$$

$$\hat{\mu}^I = \left(\hat{\mu}_x^1, \ldots, \hat{\mu}_x^1, \ldots, \hat{\mu}_x^{J_{m_1}}, \ldots, \hat{\mu}_x^{J_{m_1}}, \hat{\mu}_y^1, \ldots, \hat{\mu}_y^1, \ldots, \hat{\mu}_y^{J_{m_2}}, \ldots, \hat{\mu}_y^{J_{m_2}}\right)^T \tag{25}$$

and

$$\hat{\Sigma}^I = \frac{1}{n}\sum_{j=1}^{n}(Z_j - \hat{\mu}_I)(Z_j - \hat{\mu}_I)^T. \tag{26}$$

Here, J_{m_1}, J_{m_2} denote the number of biological replicates for X and Y, whereas $I_{m_1}^j$, $I_{m_2}^j$, $1 \leq j_{m_1} \leq J_{m_1}$, $1 \leq j_{m_2} \leq J_{m_2}$, represent the number of technical replicates nested within $j_{m_1}^{th}$ and $j_{m_2}^{th}$ biological replicate respectively, where $\sum_{j=1}^{J_{m_1}} I_{m_1}^j = m_1$ and $\sum_{j=1}^{J_{m_2}} I_{m_2}^j = m_2$. However, on averaging the off-diagonal block of $\hat{\Sigma}^I$ to estimate a single correlation value, as in the case of blind-case model (Eq. 19), between-molecular correlations from informed-case model and blind-case model become identical (see (Zhu *et al.*, 2010) for proof). To exploit the informed replication mechanism and compare model performances, likelihood ratio test based methods (Anderson, 1958) are used. Indeed, the hypothesis

$$H_0 : Z \in N(\mu, \Sigma_0) \text{ versus } H_\alpha : Z \in N(\mu, \Sigma)$$

is tested by considering $(\mu, \Sigma) = (\mu_B, \Sigma_B)$ and $(\mu, \Sigma) = (\mu_I, \Sigma_I)$. Matrix Σ_0 is obtained by setting the off-diagonal entries in Σ to 0. Likelihood ratio test statistics for blind-case and informed-case models are calculated using

$$\Psi = -2 \log(\wedge) \tag{27}$$

where

$$\wedge = \frac{|\hat{\Sigma}_0|^{-n/2} \exp(\frac{-1}{2}\sum_{j=1}^{n}(Z_j - \hat{\mu})^T \hat{\Sigma}_0^{-1}(Z_j - \hat{\mu}))}{|\hat{\Sigma}|^{-n/2} \exp(\frac{-1}{2}\sum_{j=1}^{n}(Z_j - \hat{\mu})^T \hat{\Sigma}^{-1}(Z_j - \hat{\mu}))}. \tag{28}$$

Under null hypothesis, the two statistics $\Psi^B = -2\log\wedge^B$ and $\Psi^I = -2\log\wedge^I$ corresponding to blind-case and informed-case model follow an asymptomatic chi-square distribution with 1 and $J_{m_1}J_{m_2}$ degrees of freedom, respectively. Thus, the model performances can be evaluated by comparing the P-values (P) from blind-case and informed-case models or directly comparing the difference $\Psi^I - \Psi^B$ to the chi-square distribution with $J_{m_1}J_{m_2} - 1$ degrees of freedom. For a more detailed study on informed-case model, we refer to (Zhu *et al.*, 2010).

It is clear that informed-case correlation estimator generalizes blind-case model by explicitly considering prior knowledge of experimental design. When there is only one biological replicate for each gene in replicated data, the two models become identical. Although informed-case model is useful, it is not practical to design a correlation structure that will fit for any replicated molecular profiling data. A key is to adaptively determine the underlying correlation structure by balancing between a model with a constrained set of parameters and the one without any constraints. This situation can be translated into the Expectation-Maximization (EM) framework (Dempster *et al.*, 1977), where we seek for the missing membership of a multivariate observation in either a component with a constrained set of parameters or the one with an unconstrained set of parameters. EM algorithm plays a crucial role in the following generalization of blind-case or informed-case model.

4.3 Finite mixture model

In the finite mixture model approach (Fraley & Raftery, 2002; McLachlan & Peer, 2000), density of an observation is modeled as mixture of a finite number of component densities. Such an approach can be used to shrink the correlation structure of a gene set between a constrained correlation structure and an unconstrained one. Advantages of shrinkage approach have been demonstrated in many related studies (Schäfer & Strimmer, 2005; Zhu & Hero, 2007). In the following discussion, we consider the two-component mixture model approach from (Acharya & Zhu, 2009), where the density of each multivariate observation Z_j is modeled as a mixture of two component densities denote by $f_1(Z_j)$ and $f_2(Z_j)$. This is expressed as

$$f(Z_j, \Psi) = \pi_1 \, f_1(Z_j) + \pi_2 \, f_2(Z_j), \tag{29}$$

where π_1 and π_2 stand for mixture proportions with $\pi_1 + \pi_2 = 1$ and Ψ denotes the set of all parameters in the mixture model, $j = 1, \ldots, n$. The first component in the mixture represents either blind-case or informed-case estimator, whereas the second component corresponds to the unconstrained $\sum_{i=1}^{k} m_i$-variate multivariate normal distribution. Let $\theta_i = \{\mu_i, \, \Sigma_i\}$ denote the set of parameters for the i^{th} component, $i = 1, \, 2$, where $\theta_1 = \{\mu_B, \, \Sigma_B\}$ or $\theta_1 = \{\mu_I, \, \Sigma_I\}$. Finite mixture model employs EM algorithm (McLachlan & Peer, 2000) to estimate the posterior probability that the j^{th} observation belongs to the i^{th} component of the mixture. Thus, incompleteness in the EM framework is incorporated by considering the component-indicator vectors z_j's, $j = 1, 2, \ldots, n$, where $(z_j)_i = z_{ij} = 1$ if Z_j is sampled from the i^{th} component, as unobserved. Complete data is comprised of the observations Z_j's together with the component-indicator vectors z_j's. The E step and M step at the $(k+1)^{th}$ iteration are defined as

E-step: For $i = 1, \, 2$,

$$\tau_i(Z_j; \Psi^{(k)}) = \frac{\pi_i^{(k)} f_i(Z_j; \theta_i^{(k)})}{\sum_{h=1}^{2} \pi_h^{(k)} f_h(Z_j; \theta_h^{(k)})} \tag{30}$$

where $\tau_i(Z_j; \Psi^{(k)})$ is the posterior probability that Z_j belongs to the i^{th} component.

M-step: For $i = 1, \, 2$,

$$\pi_i^{k+1} = \frac{1}{n} \sum_{j=1}^{n} \tau_i(Zj; \Psi^{(k)}) \tag{31}$$

$$\mu_i^{k+1} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} Z_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \tag{32}$$

$$\Sigma_i^{k+1} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (Z_j - \mu_i^{(k+1)})(Z_j - \mu_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \tag{33}$$

where $\tau_{ij}^{(k)} = \tau_i(Z_j; \Psi^{(k)})$. EM algorithm iterates between the E step and the M step until convergence. Finally, an observation Z_j corresponds to a component model for which it has higher posterior probability of belonging, $j = 1, 2, \ldots, n$. However, in many cases the sequence $\{\log L(\Psi^k)\}$ of log-likelihood values generated in the iterative procedure may not be bounded or it may be trapped in a local solution (McLachlan & Peer, 2000). Consequently,

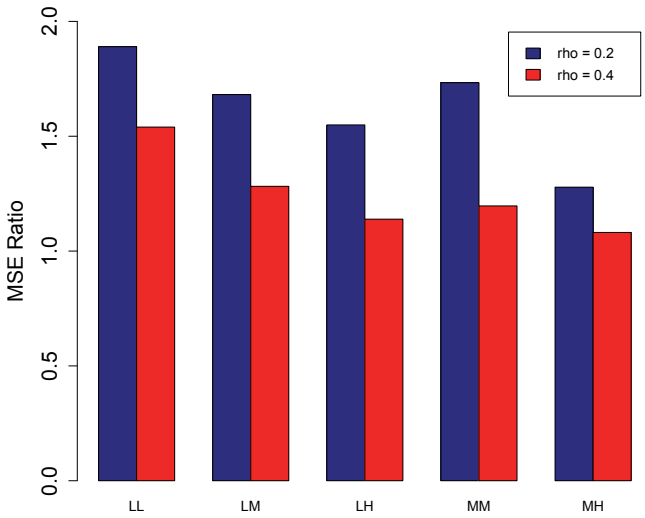

Fig. 2. Comparison of the multivariate blind-case model and bivariate Pearson's correlation estimator. In the figure, the x-axis corresponds to data quality and y-axis represents MSE ratio, which is the ratio MSE from Pearson's estimator/MSE from blind-case model. Pair of genes, each with 4 replicated measurements across 20 samples, were considered in the comparison. The between molecular correlation parameter (rho) was set at 0.2 (low) and 0.4 (medium), respectively.

the unconstrained EM algorithm presented above may not necessarily converge to the MLE $\hat{\Psi}$. To reduce various problems associated with the convergence of EM algorithm, remedies have been proposed by constraining the eigenvalues of the component correlation matrices (Ingrassia, 2004; Ingrassia & Rocci, 2007). For example, the constrained EM algorithm presented in (Ingrassia, 2004) considers two strictly positive constants a and b such that $a/b \geq c$, where $c \in (0\ 1]$. In each iteration of the EM algorithm, if the eigenvalues of the component correlation matrices are smaller than a, they are replaced with a and if they greater than b, they are replaced with b. Indeed, if the eigenvalues of the component correlation matrices satisfy $a \leq \lambda_j(\Sigma_i) \leq b$, for $i = 1, 2, j = 1, 2, \ldots, \sum_{i=1}^{k} m_i$, then the condition $\lambda_{\min}(\Sigma_1 \Sigma_2^{-1}) \geq c$ (Hathaway, 1985) is also satisfied, and results in constrained (global) maximization of the likelihood.

5. Results

5.1 Simulations

In this section, we evaluate the performance of multivariate and bivariate correlation estimators using synthetic replicated data. In Figure 2, we compare multivariate blind-case model and bivariate Pearson's correlation estimator by simulating 1000 synthetic data sets corresponding to a pair of genes, each with 4 replicated measurements and 20 observations.

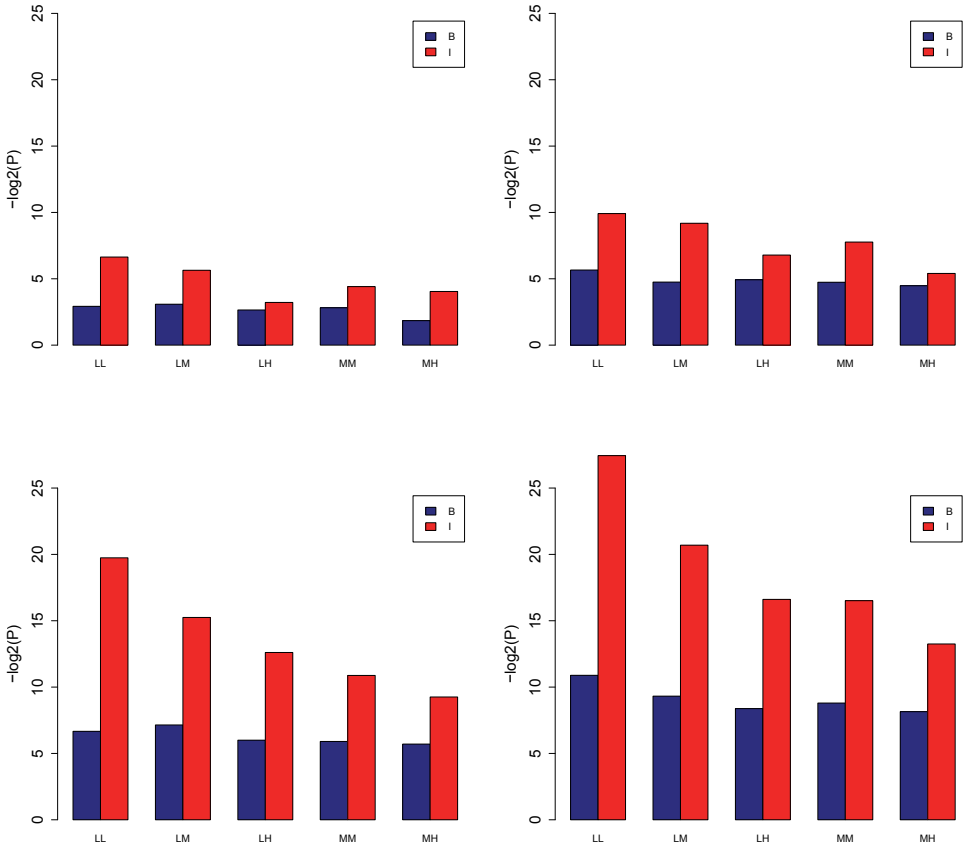

Fig. 3. Comparison of the multivariate blind-case model and informed-case model with increasing data quality and sample size, as presented in (Zhu *et al.*, 2010). Pair of genes, each with 3 biological replicates and 2 technical replicates nested within a biological replicate, were considered in the comparison. The range of between-molecular correlation parameters was set at M (0.3-0.5). Two upper panels correspond to replicated data with sample size $n = 20$ (left) and $n = 30$ (right), and the lower panels correspond to the ones with $n = 40$ (left) and $n = 50$ (right).

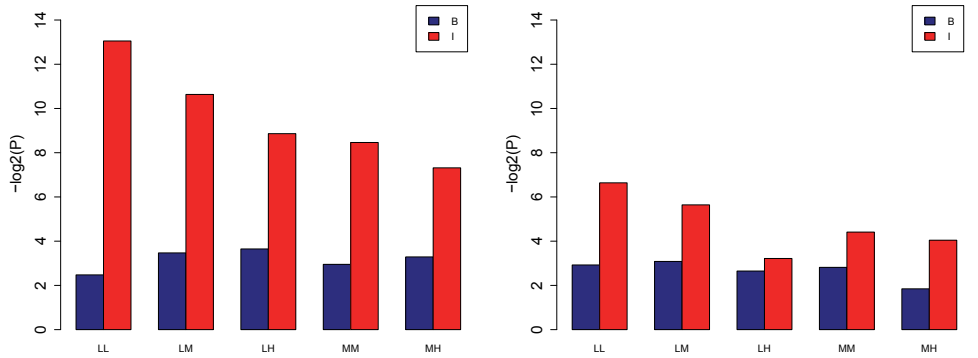

Fig. 4. Comparison of the multivariate blind-case model and informed-case model with increasing number of technical replicates, as presented in (Zhu *et al.*, 2010). Pair of genes, each with 3 biological replicates and 20 observations were considered in the comparison. The range of between-molecular correlation parameters was set at M (0.3-0.5). The left and right panels correspond to 1 and 2 technical replicates nested within a biological replicate, respectively.

Along the x-axis, L (low: $0.1 - 0.3$), M (medium: $0.3 - 0.5$) and H (high: $0.5 - 0.7$) represent the range of within-molecular correlations for each of the two genes. The y-axis corresponds to MSE (mean squared error) ratio, which is the ratio of MSE from Pearson's estimator over MSE from blind-case model. Thus, MSE ratio greater than 1 indicates the superior performance of blind-case model. We fixed the between molecular correlation parameter at 0.2 (low) and 0.4 (medium), respectively. As shown in Fig. 2, all examined MSE ratios were found greater than 1. Figure 2 also demonstrates that the performance of blind-case model is a decreasing function of data quality. This observation makes blind-case model particularly suitable for analyzing real-world replicated data sets, which are often contaminated with excessive noise. Figure 3 and Figure 4 represent parts of more detailed studies conducted in (Zhu *et al.*, 2010) to evaluate the performances of multivariate correlation estimators. For instance, Figure 3 compares the multivariate blind-case model and informed-case model with increasing data quality and sample size. Synthetic data sets corresponding to a pair of genes, each with 3 biological replicates and 2 technical replicates nested within a biological replicate in 20 experiments were used in the comparison. The model performances were estimated in terms of $-\log_2(P)$ values. Higher $-\log_2(P)$ values indicate better performance by a model. As demonstrated in Fig. 3, informed-case model significantly outperformed the blind-case model in estimating pairwise correlation from replicated data with informed replication mechanisms. It is also observed in Figure 3 that blind-case and informed-case models are increasing functions of sample size and decreasing functions of data quality. The two models were also compared in terms of increasing number of technical replicates of a biological replicate, as demonstrated in Figure 4. We conclude from Figure 4 that blind-case and informed-case models are decreasing functions of the number of technical replicates nested with a biological replicate.

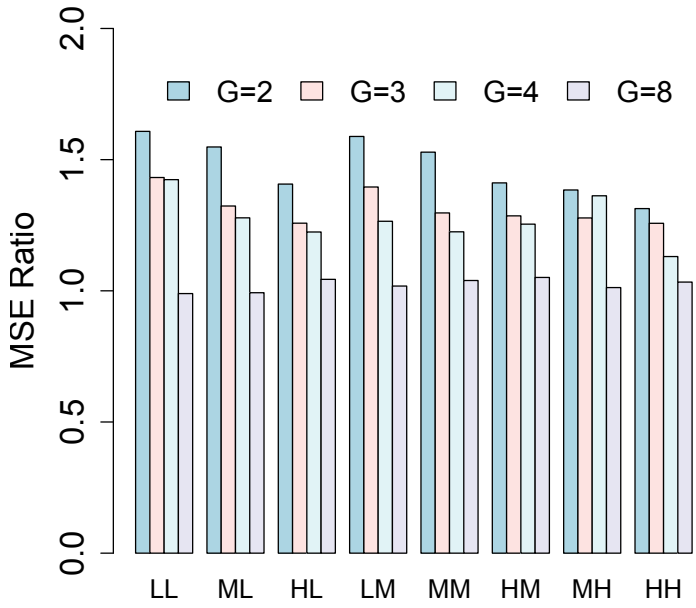

Fig. 5. Comparison of the multivariate blind-case model and two-component finite mixture model in terms of MSE ratio, as presented in (Acharya & Zhu, 2009). MSE ratio is calculated as MSE from blind-case model/MSE from mixture model. Gene sets with 2, 3, 4 and 8 genes, each with 4 replicated measurements across 20 samples were considered in the comparison.

Fig. 5, originally from (Acharya & Zhu, 2009), compares the performance of blind-case model and two component finite mixture model in estimating the correlation structure of a gene set. The constrained component in the mixture model corresponds to blind-case correlation estimator. Fig. 5 plots the model performances in terms of MSE ratio defined as MSE from blind-case model/MSE from mixture model. The number of genes in a gene set are fixed at $G = 2,\ 3,\ 4$ and 8. In Fig. 5, almost all examined MSE ratios greater than 1 indicate an overall better performance of the mixture model approach compared with blind-case model. Fig. 5 also indicates that the performance of finite mixture model is a decreasing functions of data quality and number of genes in the input.

5.2 Real-world data analysis

In Figure 6-8, we present real-world studies conducted in (Acharya & Zhu, 2009), where blind-case model and finite mixture model were used to analyze two publically available replicated data sets, spike-in data from Affymetrix (`http://www.affymetrix.com`) and yeast galactose data (`http://expression.washington.edu/publications/kayee`) from (Yeung *et al.*, 2003). Spike-in data comprises of the gene expression levels of 16 genes

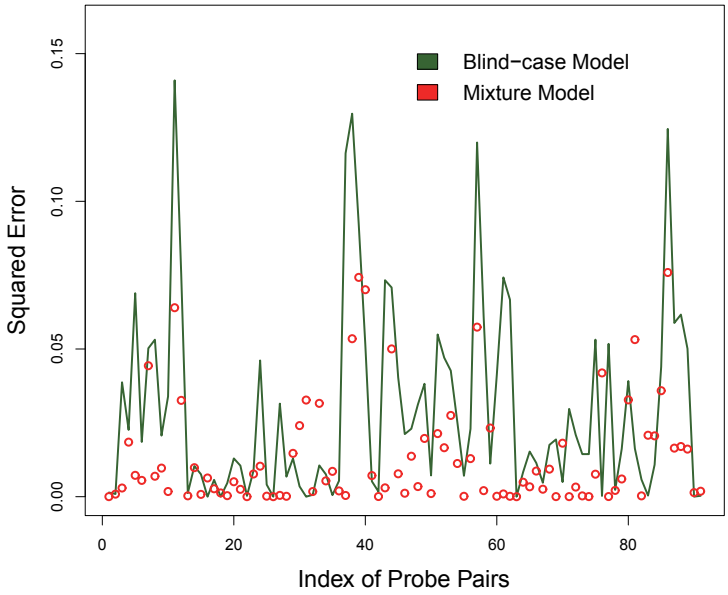

Fig. 6. Comparison of two multivariate models, blind-case model and finite mixture model, in estimating pairwise correlations among genes in spike-in data, as presented in (Acharya & Zhu, 2009).

in 20 experiments, where 16 replicated measurements are available for a gene. Correlation structures estimated using spike-in data were compared with the nominal correlation structure obtained from a prior known probe-level intensities. On the other hand, yeast data contains the gene expression levels of 205 genes, each with 4 replicated measurements. Yeast data was used to assess model performances in hierarchial clustering by utilizing a prior knowledge of the class labels of 205 genes.

Figure 6 compares the performance of blind-case model and mixture model in estimating pairwise correlation between genes present in spike-in data. We observed that for almost 82% of the probe pairs, mixture model provided a better approximation to the nominal pairwise correlation compared with blind-case model. The two models were further employed to estimate the correlation structure of a gene set. Figure 7 corresponds to the correlation structure of a collection of 10 randomly selected probe sets from spike-in data. As demonstrated in Figure 7, an overall better performance of mixture model approach was given by lower squared error in comparison to blind-case model.

Finally, blind-case model and mixture model were utilized to estimate the correlation structures from 150 subsets of yeast data, each with 60 randomly selected probe sets. The estimated correlation structures were used to perform correlation based hierarchial clustering. Figure 8 compares the clustering performance of blind-case model and mixture model in terms of Minkowski score. Minkowski score is defined as $\|C - T\|/\|T\|$, where C and T are binary matrices constructed from the predicted and true labels of genes, respectively. C_{ij}

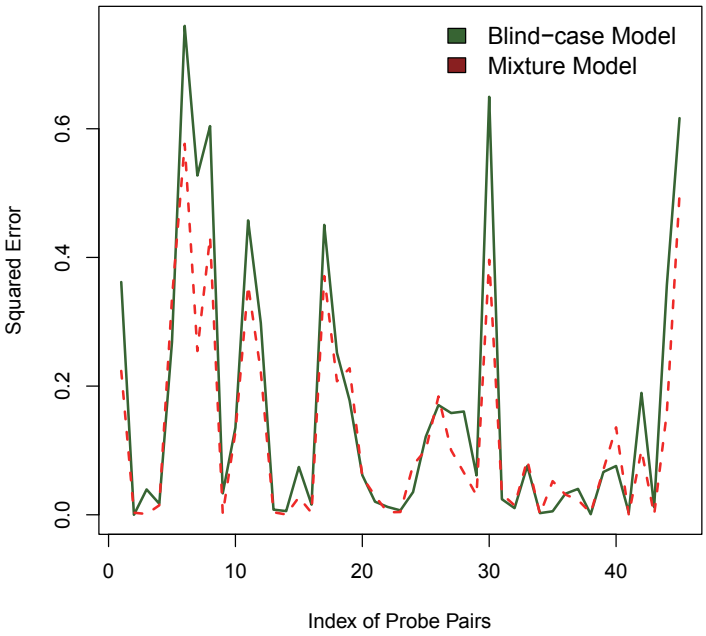

Fig. 7. Comparison of the multivariate blind-case model and finite mixture model in estimating the correlation structure of a gene set, as presented in (Acharya & Zhu, 2009). The figure corresponds to a gene set comprising of 10 randomly selected probe sets in spike-in data. Each index along the x-axis represents a probe set pair and y-axis plots squared error values in estimating nominal correlations.

=1, if i^{th} and j^{th} gene belong to the same cluster in the solution and 0 otherwise. Matrix T is obtained analogously using the true labels. A lower Minkowski score indicates higher clustering accuracy. In Figure 8, an overall better performance of two-component mixture model approach was observed in almost 73% cases.

6. Conclusions

Rapid developments in high-throughput data acquisition technologies have generated vast amounts of molecular profiling data which continue to accumulate in public databases. Since such data are often contaminated with excessive noise, they are replicated for a reliable pattern discovery. An accurate estimate of the correlation structure underlying replicated data can provide deep insights into the complex biomolecular activities. However, traditional bivariate approaches to correlation estimation do not automatically accommodate replicated measurements. Typically, an *ad hoc* step of data preprocessing by averaging (weighted, unweighted or something in between) is needed. Averaging creates a strong bias while reducing variance among the replicates with diverse magnitudes. It may also wipe out

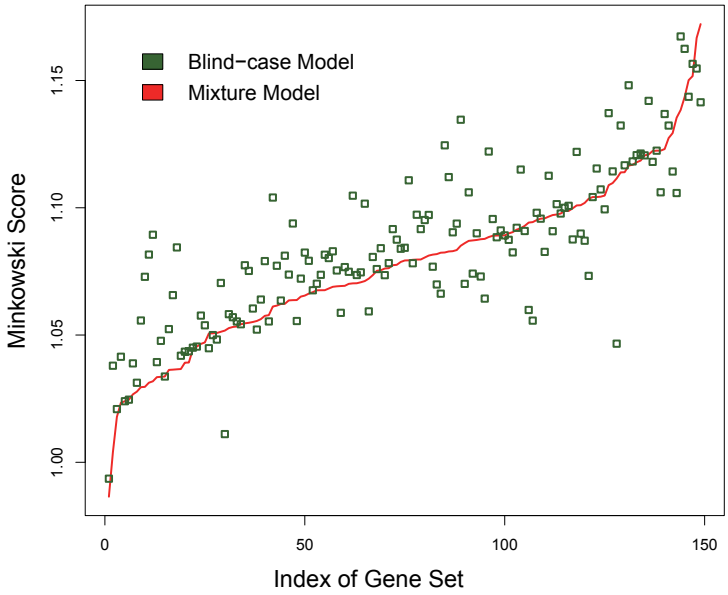

Fig. 8. Performance of the multivariate blind-case model and finite mixture model in clustering yeast data, as presented in (Acharya & Zhu, 2009). Each index along the x-axis corresponds to a subset of yeast data comprising of 60 randomly selected probe sets. The y-axis plots model performances in terms of Minkowski score. An overall better performance of the mixture model approach is given by lower Minkowski scores in almost 73% cases.

important patterns of small magnitudes or cancel out patterns of similar magnitudes. In many cases prior knowledge of the underlying replication mechanism might be known. However, this information can not be exploited by averaging replicated measurements. Thus, it is necessary to design multivariate approaches by treating each replicate as a variable. In this chapter, we reviewed two bivariate correlation estimators, Pearson's correlation and SD-weighted correlation, and three multivariate models, blind-case model, informed-case model and finite mixture model to estimate the correlation structure from replicated molecular profiling data corresponding to a gene set with blind or informed replication mechanism. Each of the three multivariate models treat a replicated measurement individually as a random variable by assuming that data as independently and identically distributed samples from a multivariate normal distribution. Blind-case model utilizes a constrained set of parameters to define the correlation structure of a gene set with blind replication mechanism, whereas informed-case model generalizes blind-case model by incorporating prior knowledge of experimental design. Finite mixture model presents a more general approach of shrinking between a constrained model, either blind-case model or informed-case model, and the unconstrained model. The aforementioned multivariate models were used to analyze synthetic and real-world replicated data sets. In practice, the choice of a multivariate correlation estimator may depend on various factors, e.g. number of genes, number of

replicated measurements available for a gene, prior knowledge of experimental design etc. For instance, blind-case and informed-case models are more stable and computationally more efficient than iterative EM based finite mixture model approach. However, considering the real-world scenarios, finite mixture model assumes a more faithful representation of the underlying correlation structure. Nonetheless, the multivariate models presented here are sufficiently generalized to incorporate both blind and informed replication mechanisms, and open new avenues for future supervised and unsupervised bioinformatics researches that require accurate estimation of correlation, e.g. gene clustering, gene networking and classification problems.

7. References

Acharya LR and Zhu D (2009). Estimating an Optimal Correlation Structure from Replicated Molecular Profiling Data Using Finite Mixture Models. In the Proceedings of *IEEE International Conference on Machine Learning and Applications*, 119-124.

Altay G and Emmert-Streib F (2010). Revealing differences in gene network inference algorithms on the network-level by ensemble methods. *Bioinformatics*, 26(14), 1738-1744.

Anderson TW (1958). *An introduction to mutilvariate statistical analysis*, Wiley Publisher, New York.

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R and Califano, A (2005). Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37:382-390.

Boscolo R, Liao J, Roychowdhury VP (2008). An Information Theoretic Exploratory Method for Learning Patterns of Conditional Gene Coexpression from Microarray Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15-24.

Butte AJ and Kohane IS (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 5, 415-426.

Casella G and Berger RL (1990). *Statistical inference*, Duxbury Advanced Series.

Dempster AP, Laird NM and Rubin DB (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1-38.

Eisen M, Spellman P, Brown PO, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95:14863-14868.

Fraley C and Raftery AE (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611-631.

Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, Che D, Dickinson T, Wickham E, Bierle J, Doucet D, Milewski M, Yang R, Siegmund C, Haas J, Zhou L, Oliphant A, Fan JB, Barnard S and Chee MS (2004). Decoding randomly ordered DNA arrays. *Genome Research*, 14:870-877.

Hastie T, Tibshirani R and Friedman J (2009). *The Elements of Statistical Learning: Prediction, Inference and Data Mining*, Springer-Verlag, New York.

Hathaway RJ (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, **13**, 795-800.

de Hoon MJL, Imoto S, Nolan J and Miyano S (2004). Open source clustering software. *Bioinformatics*, 20(9):1453-1454.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H and He YD (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102:109-126.

Ingrassia S (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications*, 13, 151-166.

Ingrassia S and Rocci R (2007). Constrained monotone EM algorithms for the finite mixtures of multivariate Gaussians. *Computational Statistics and Data Analysis*, 51, 5399-5351.

Kerr MK and Churchill GA (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2:183-201.

Kung C, Kenski DM, Dickerson SH, Howson RW, Kuyper LF, Madhani HD, Shokat KM (2005). Chemical genomic profiling to identify intracellular targets of a multiplex kinase inhibitor. *Proceedings of the National Academy of Sciences*, 102:3587-3592.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H and Brown EL (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675-1680.

McLachlan GJ and Peel D (2000). *Finite Mixture Models*. Wiley series in Probability and Mathematical Statistics, John Wiley & Sons.

McLachlan GJ and Peel D (2000). On computational aspects of clustering via mixtures of normal and t-components. *Proceedings of the American Statistical Association*, Bayesian Statistical Science Section, Indianapolis, Virginia.

Medvedovic M and Sivaganesan S (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18:1194-1206.

Medvedovic M, Yeung KY and Bumgarner RE (2004). Bayesian mixtures for clustering replicated microarray data. *Bioinformatics*, 20:1222-1232.

Rengarajan J, Bloom BR and Rubin EJ (2005). From The Cover: Genomewide requirements for Mycobacterium tuberculosis adaptation and survival in macrophages. *Proceedings of the National Academy of Sciences*, 102(23):8327-8332.

Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD and Medvedovic, M (2006) Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics*, 7:538.

Schäfer J and Strimmer K (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4, Article 32.

Shendure J and Ji H (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26, 1135-1145.

van't Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530-536.

Yao J, Chang C, Salmi ML, Hung YS, Loraine A and Roux SJ (2008). Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient. *BMC Bioinformatics*, 9:288.

Yeung KY, Medvedovic M and Bumgarner R. (2003). Clustering gene expression data with repeated measurements. *Genome Biology*, 4:R34.

Yeung KY and Bumgarner R (2005). Multi-class classification of microarray data with repeated measurements: application to cancer. *Genome Biology*, 6(405).

Zhu D, Hero AO, Qin ZS and Swaroop A (2005). High throughput screening co-expressed gene pairs with controlled biological significance and statistical significance. *Journal of Computational Biology*, 12(7):1029-1045.

Zhu D, Li Y and Li H (2007). Multivariate correlation estimator for inferring functional relationships from replicated genome-wide data. *Bioinformatics*, 23(17):2298-2305.

Zhu D and Hero AO (2007). Bayesian hierarchical model for large-scale covariance matrix estimation. *Journal of Computational Biology*, 14(10):1311-1326.

Zhu D, Acharya LR and Zhang H (2010). A Generalized Multivariate Approach to Pattern Discovery from Replicated and Incomplete Genome-wide Measurements, *IEEE/ACM transaction on Computational Biology and Bioinformatics*, (in press).

Advanced Biomedical Engineering

Edited by Dr. Gaetano Gargiulo

ISBN 978-953-307-555-6
Hard cover, 280 pages
Publisher InTech
Published online 23, August, 2011
Published in print edition August, 2011

This book presents a collection of recent and extended academic works in selected topics of biomedical signal processing, bio-imaging and biomedical ethics and legislation. This wide range of topics provide a valuable update to researchers in the multidisciplinary area of biomedical engineering and an interesting introduction for engineers new to the area. The techniques covered include modelling, experimentation and discussion with the application areas ranging from acoustics to oncology, health education and cardiovascular disease.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Lipi R. Acharya and Dongxiao Zhu (2011). Multivariate Models and Algorithms for Learning Correlation Structures from Replicated Molecular Profiling Data, Advanced Biomedical Engineering, Dr. Gaetano Gargiulo (Ed.), ISBN: 978-953-307-555-6, InTech, Available from: http://www.intechopen.com/books/advanced-biomedical-engineering/multivariate-models-and-algorithms-for-learning-correlation-structures-from-replicated-molecular-pro

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821